Anticipating Visual Representations from Unlabeled Data

Carl Vondrick, Hamed Pirsiavash, Antonio Torralba

Overview

- **Problem**
- Key Insight
- Methods
- Experiments

Problem: Predict future actions and objects

Image from Vondrick et. al, http://web.mit.edu/vondrick/prediction/paper.pdf

Related Work

- Unlabeled video prediction
  - Motion and trajectory prediction
  - Pixel level prediction
- Action prediction
  - Intention inference
  - Semantic context for action prediction
- Path and motion prediction
  - Optical flow

Applications

- Robotics
  - Path planning
  - Human robot interaction
  - Obstacle avoidance
- Surveillance
  - Warning systems

Overview

- Problem
- **Key Insight**
- Methods
- Experiments

Key Insight: Don’t predict images

Image from Vondrick et. al, http://web.mit.edu/vondrick/prediction/paper.pdf

Key Insight: Predict Intermediate Representation

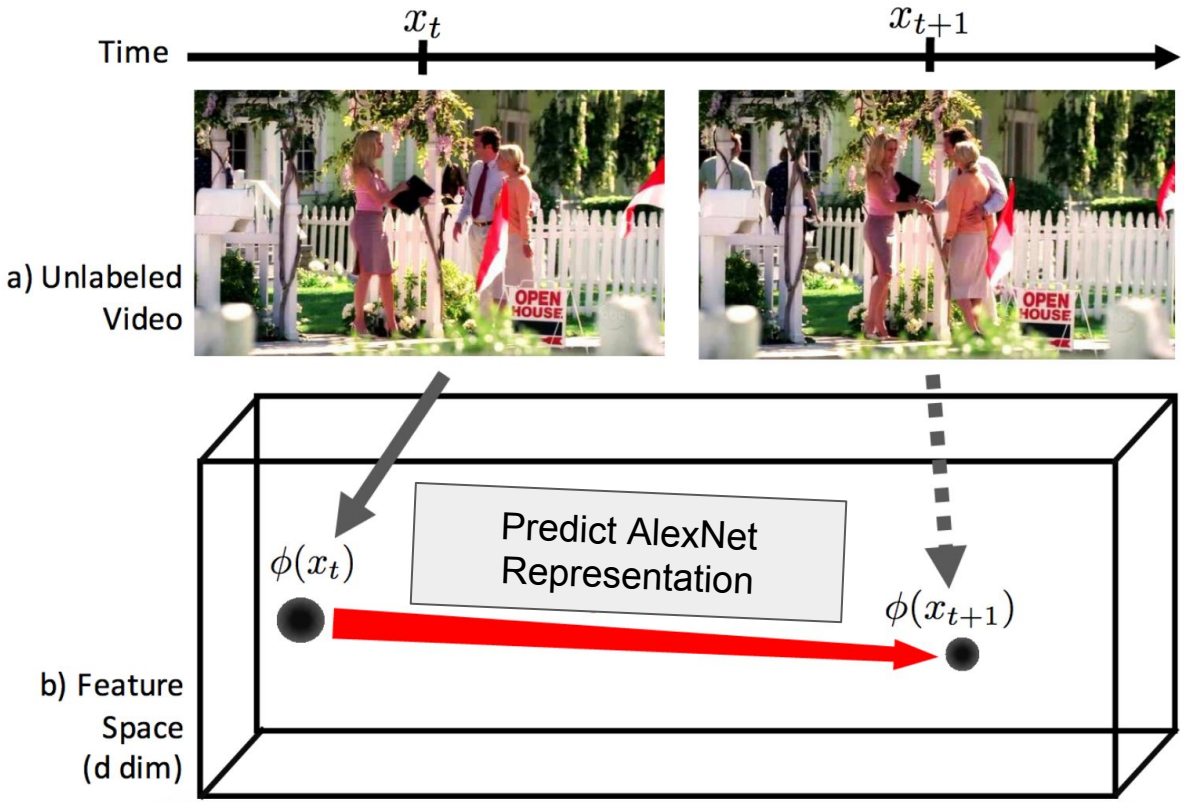


Image from Vondrick et. al, http://web.mit.edu/vondrick/prediction/paper.pdf

Key Insight: Predict Intermediate Representation

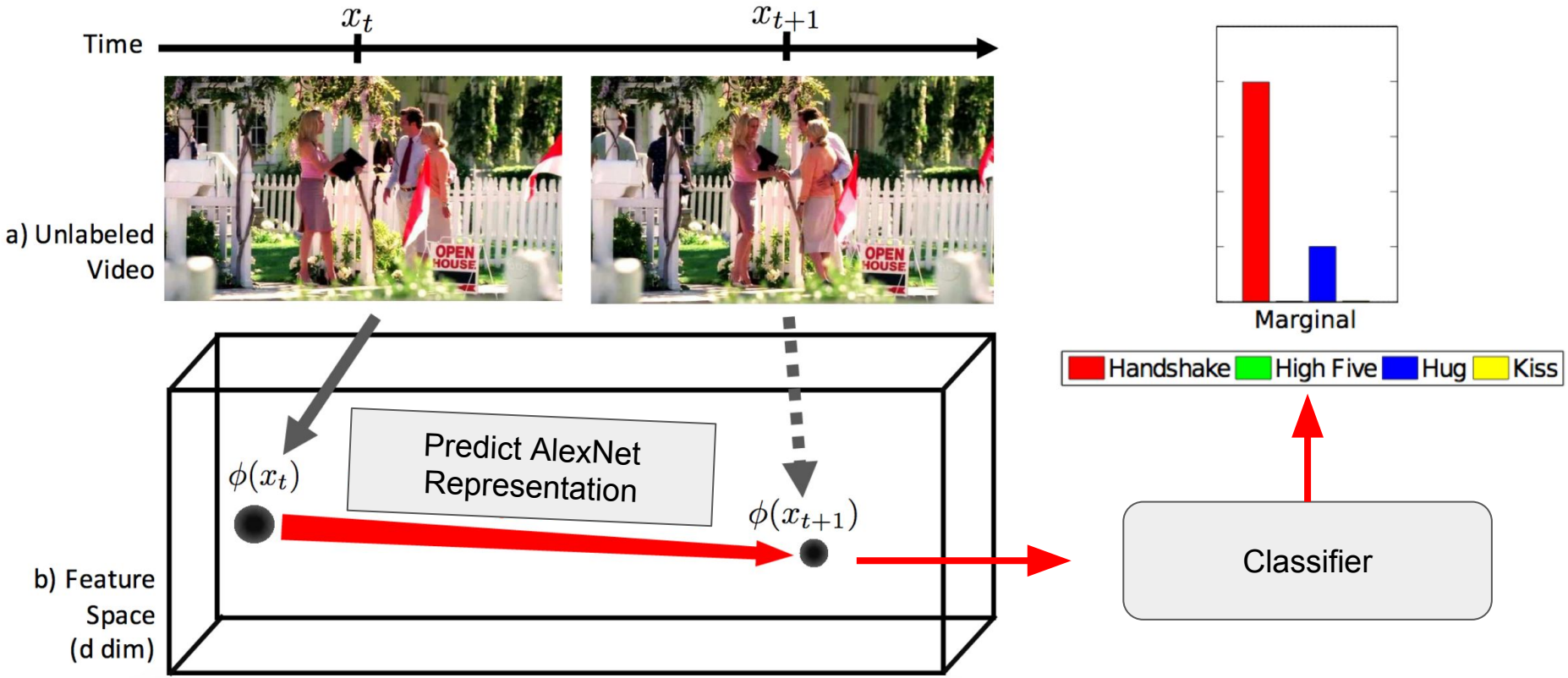


Image from Vondrick et. al, http://web.mit.edu/vondrick/prediction/paper.pdf

Overview

- Problem
- Key Insight
- **Methods**
- Experiments

Use Unlabeled Video as Training Data

- The internet is full of unlabeled videos
  - Used 600 hours of popular TV shows on YouTube
- Get supervision for free!
  - (Because they all go forward in time)
- Can then use predicted representation for action or object detection

Multiple Futures

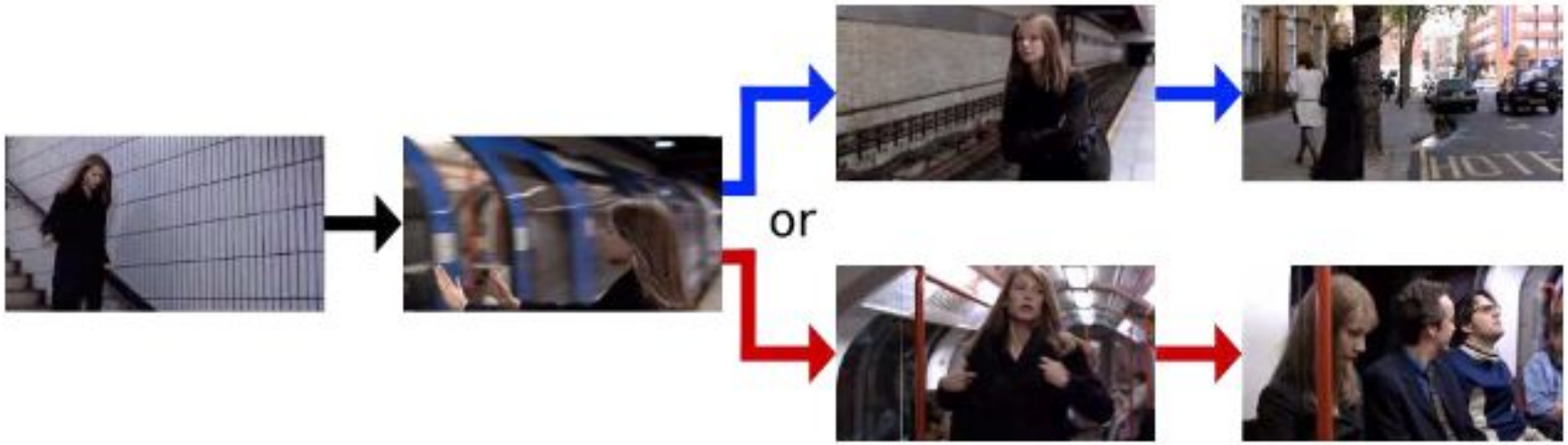
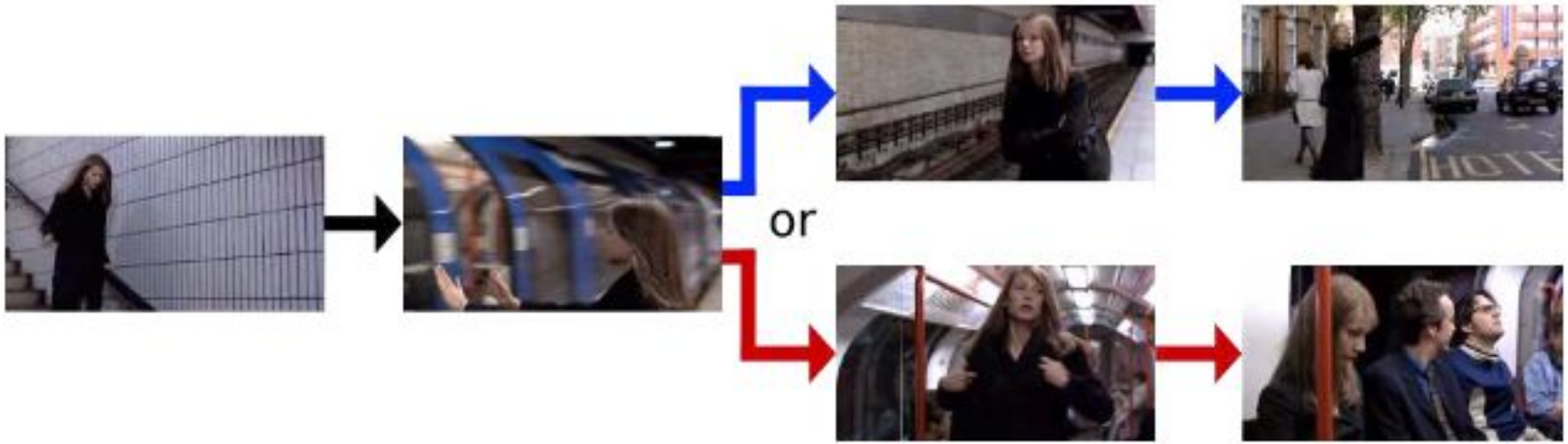


Image from Vondrick et. al, http://web.mit.edu/vondrick/prediction/paper.pdf

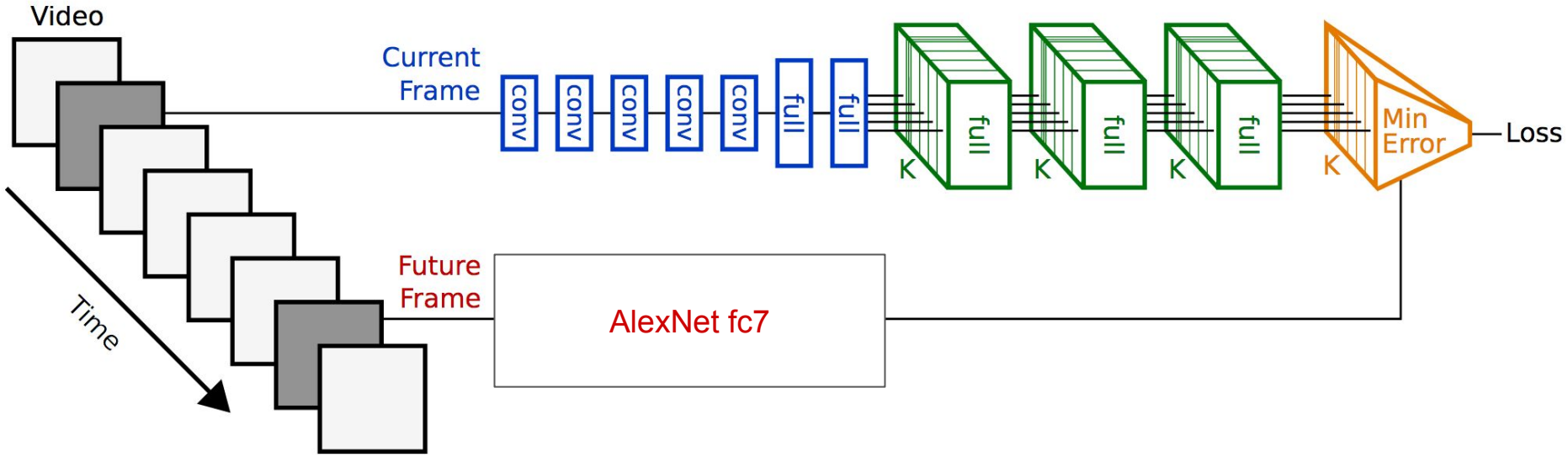
Multiple Futures



- Train network to predict K representations for the future
- Classify all K representations
- Predict class with highest marginal probability

Image from Vondrick et. al, http://web.mit.edu/vondrick/prediction/paper.pdf

Network Architecture



- Alexnet with additional fully connected layers
- Loss function is simply argmin of squared error.

Image from Vondrick et. al, http://web.mit.edu/vondrick/prediction/paper.pdf

Overview

- Problem
- Key Insight
- Methods
- **Experiments**

Action Forecasting Experiment

- Dataset: TV Human Interactions
  - 300 separate videos
  - People do one of: {high fiving, hugging, shaking hands, kissing}
- Goal: Predict activity 1 second in the future
- Baselines:
  - SVM, Nearest Neighbor, Max Margin Event Detector, Linear Regression

Normal vs. Adapted Training: Normal Training

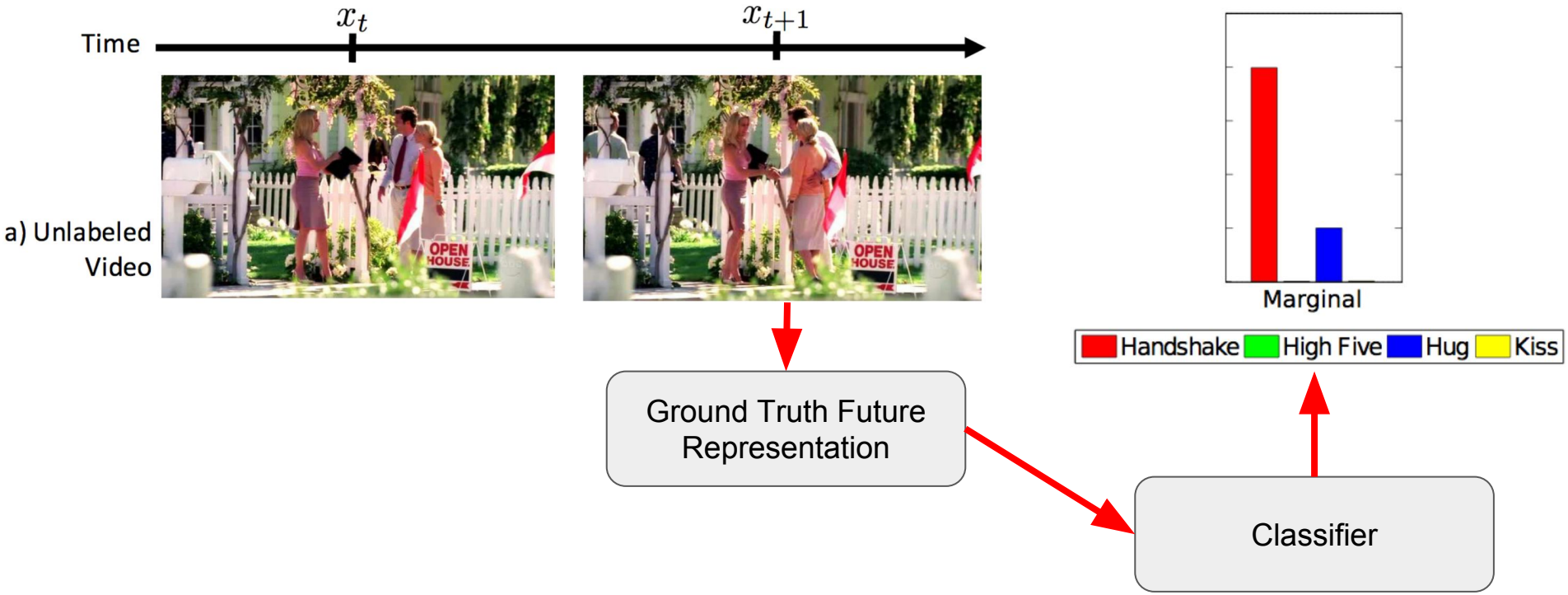


Image from Vondrick et. al, http://web.mit.edu/vondrick/prediction/paper.pdf

Normal vs. Adapted Training: Adapted Training

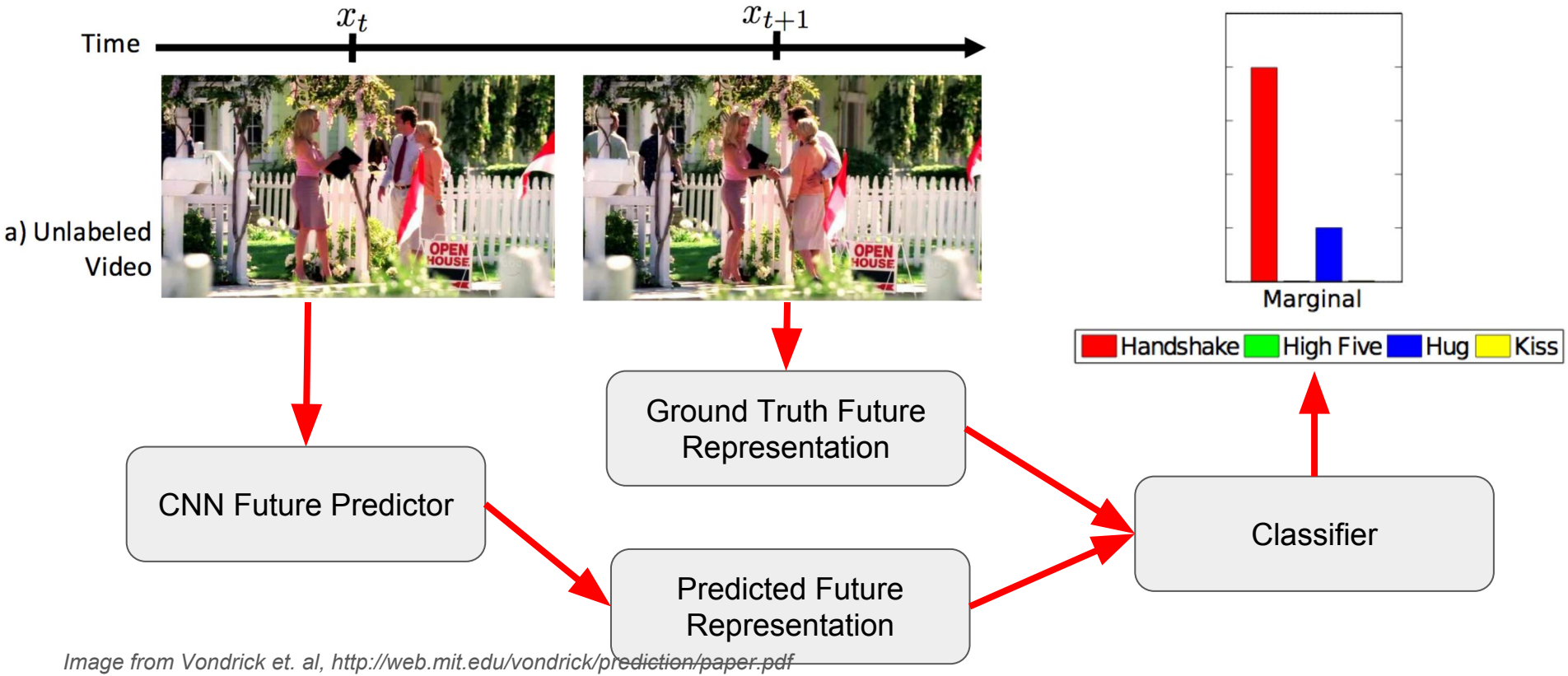


Image from Vondrick et. al, http://web.mit.edu/vondrick/prediction/paper.pdf

Action Forecasting Results

Method	Accuracy
Random	25.0
SVM Static	36.2 ± 4.9
SVM	35.8 ± 4.3
MMED	34.0 ± 7.0
Nearest Neighbor	29.9 ± 4.6
Nearest Neighbor [43], Adapted	34.9 ± 4.7
Linear	32.8 ± 6.1
Linear, Adapted	34.1 ± 4.8
Deep K=1, ActionBank [34]	34.0 ± 6.1
Deep K=3, ActionBank [34]	35.7 ± 6.2
Deep K=1	36.1 ± 6.4
Deep K=1, Adapted	40.0 ± 4.9
Deep K=3	35.4 ± 5.2
Deep K=3, Adapted	$\mathbf{43.3 \pm 4.7}$
Deep K=3, THUMOS [9], Off-the-shelf	29.1 ± 3.9
Deep K=3, THUMOS [9], Adapted	$\mathbf{43.6 \pm 4.8}$
Human (single)	71.7 ± 4.2
Human (majority vote)	85.8 ± 1.6

- Deep, adapted networks outperform all baselines
- Much effort needed to approach human-level performance

Action Forecasting Results

Action Forecasting Results?

Object Forecasting Experiment

- Dataset: Daily Living Activities Dataset
  - Egocentric video
  - Segments featuring 1 of 14 objects
- Goal: Predict object on screen 5 seconds in the future
- Baselines:
  - SVM, Scene features, Linear Classifier
- Normal & Adapted, as before

Object Forecasting Results

Method	Mean	dish	door	utensil	cup	oven	person	soap	tap	tbrush	tpaste	towel	trashc	tv	remote
Random	1.2	1.2	2.8	1.1	2.4	1.6	0.8	1.5	2.1	0.2	0.3	0.6	1.1	0.5	0.3
SVM Static	6.4	2.6	15.4	2.9	5.0	9.4	6.9	11.5	17.6	1.6	1.0	1.5	6.0	2.0	5.9
SVM	5.3	3.0	8.2	5.2	3.6	8.3	12.0	6.7	11.7	3.5	1.5	4.9	1.3	0.9	4.1
Scene	8.2	3.3	18.5	5.6	3.6	18.2	10.8	9.2	6.8	8.0	8.1	5.1	5.7	2.0	10.3
Scene, Adapted	7.5	4.6	9.1	6.1	5.7	15.4	13.9	5.0	15.7	13.6	3.7	6.5	2.4	1.8	1.7
Linear	6.3	7.5	9.3	7.2	5.9	2.8	1.6	13.6	15.2	3.9	5.6	2.2	2.9	2.3	7.8
Linear, Adapted	5.3	2.8	13.5	3.8	3.6	11.5	11.2	5.8	4.9	5.4	3.3	3.4	1.6	2.1	1.0
Deep K=1	9.1	4.4	17.9	3.0	14.8	11.9	9.6	17.7	15.1	6.3	6.9	5.0	5.0	1.3	8.8
Deep K=1, Adapted	8.7	3.5	11.0	9.0	6.5	16.7	16.4	8.4	22.2	12.4	7.4	5.0	1.9	1.6	0.5
Deep K=3	**10.7**	4.1	22.2	5.7	16.4	17.5	8.4	19.5	20.6	9.2	5.3	5.6	4.2	8.0	2.6
Deep K=3, Adapted	**10.1**	3.5	14.7	14.2	6.7	14.9	15.8	8.6	29.7	12.6	4.6	10.9	1.8	1.4	1.9

- Performance indicates that this is a difficult task
  - Still, outperformed all other methods.

Object Forecasting Results

Image from Vondrick et. al, http://web.mit.edu/vondrick/prediction/paper.pdf

Future Work

- More robust experimentation
  - Comparison to other action prediction systems
  - Improved performance on egocentric dataset
  - Datasets (i.e. THUMOS) where semantic roles are not implicit
- Extension to real world problems
  - Robotics, surveillance, etc.
- Video Generation

Conclusion

- Problem
  - Predicting future actions or objects in video
- Key Insight
  - Learn to predict intermediate representations from unlabeled data
- Methods
  - AlexNet with additional FC layers
- Experiments
  - Outperformed baselines on action detection, still work to do to reach human performance
  - Object forecasting results proved to be challenging, still outperformed baselines

The End.